

International Journal of Modern Physics C
 © World Scientific Publishing Company

COMPLEX NETWORK ANALYSIS OF LITERARY AND SCIENTIFIC TEXTS

IWONA GRABSKA-GRADZIŃSKA

*Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University,
 ul. Reymonta 4, 30-059 Kraków, Poland
 grabska@gmail.com*

ANDRZEJ KULIG

*Institute of Nuclear Physics, Polish Academy of Sciences,
 ul. Radzikowskiego 152, 31-342 Kraków, Poland
 andrzej.kulig@ifj.edu.pl*

JAROSŁAW KWAPIEŃ

*Institute of Nuclear Physics, Polish Academy of Sciences,
 ul. Radzikowskiego 152, 31-342 Kraków, Poland
 jaroslaw.kwapien@ifj.edu.pl*

STANISŁAW DROŻDŹ

*Institute of Nuclear Physics, Polish Academy of Sciences,
 ul. Radzikowskiego 152, 31-342 Kraków, Poland
 Faculty of Physics, Mathematics and Computer Science, Cracow University of Technology,
 ul. Warszawska 24, 31-155 Kraków, Poland
 stanislaw.drozdz@ifj.edu.pl*

Received Day Month Year
 Revised Day Month Year

We present results from our quantitative study of statistical and network properties of literary and scientific texts written in two languages: English and Polish. We show that Polish texts are described by the Zipf law with the scaling exponent smaller than the one for the English language. We also show that the scientific texts are typically characterized by the rank-frequency plots with relatively short range of power-law behavior as compared to the literary texts. We then transform the texts into their word-adjacency network representations and find another difference between the languages. For the majority of the literary texts in both languages, the corresponding networks revealed the scale-free structure, while this was not always the case for the scientific texts. However, all the network representations of texts were hierarchical. We do not observe any qualitative and quantitative difference between the languages. However, if we look at other network statistics like the clustering coefficient and the average shortest path length, the English texts occur to possess more clustered structure than do the Polish ones. This result was attributed to differences in grammar of both languages, which was also indicated in the Zipf plots. All the texts, however, show network structure that differs from

2 Authors' Names

any of the Watts-Strogatz, the Barabasi-Albert, and the Erdos-Renyi architectures.

Keywords: Complex networks; Zipf law; Natural language

PACS Nos.: 64.60aq, 89.75Fb

1. Introduction

Natural language is an evolving system whose present structure can doubtlessly be considered a product of long history of self-organization¹. Like for many other self-organized systems known in Nature, the observables associated with language, being, for example, written texts or spoken messages, reveal quite sophisticated dynamics. Any language sample by no means is an amorphous mixture of symbols (letters, phonemes, morphemes, words, etc.) but rather a highly organized sequence in which particular symbols are ordered according to specific rules most of which are defined by the language grammar. Since the existence of grammar is an emergent phenomenon^{2,3}, language can be counted among the complex systems^{4,5}. The grammatical rules together with the information content impose on the language elements relations which can be most easily expressed in a form of network where, for instance, words are expressed by nodes and their relations by edges. Some earlier attempts along this way were presented in Refs.^{6,7,8,9,10} for English, Portuguese and Chinese. Here we show a few results that were obtained for English and Polish and for different types of texts (literary or scientific).

2. Methods and data

Our analysis was based on texts samples written in two languages: English and Polish. Both belong to the Indo-European family, but to different groups: the West-Germanic and the West-Slavic group, respectively. Their grammar therefore significantly differs, most notably in the existence of rich inflection of words in Polish as compared to a rather residual one in English. However, in the present work we do not deal with the semantic analysis, but restrict our study to a statistical analysis of word adjacency. As regards the English part, we analyze two groups of texts. The first one comprises the literary texts represented by 9 works of prose ("Ulysses" and "Finnegans Wake" by J. Joyce, "Alice's Adventures in Wonderland" by L. Carroll, "Adventures of Huckleberry Finn" by M. Twain, "Pride and Prejudice" by J. Austen, "Oliver Twist" by C. Dickens, "Secret Adversary" by A. Christie, "Adventures of Sherlock Holmes" and "Study in Scarlet" by A. Conan Doyle), 4 dramas by W. Shakespeare ("Hamlet", "Macbeth", "Winter Tale", and "Romeo and Juliet"), 61 poems by O. Wilde, and 25 poems by T.S. Elliott. The second group comprises the scientific texts represented by selected works of E. Witten¹¹, E.G.D. Cohen¹², S. Weinberg¹³, and P.W. Anderson¹⁴, as well as by three long reviews by D. Sornette¹⁵, R. Albert and A.-L. Barabasi¹⁶, and J. Kwapień and S. Drożdż⁵. Somewhere at the interface of these two groups, there is "A Brief History of Time" by S. Hawking and "The Emperor's New Mind: Con-

cerning Computers, Minds, and the Laws of Physics” by R. Penrose representing popular science. The Polish language was represented by the novels: “Lalka” (“The Doll”) by B. Prus, “Bramy Raju” (“The Gates of Paradise”) by J. Andrzejewski, “Dolina Issy” (“The Issa Valley”) by C. Miłosz, “Cesarz” (“The Emperor: Downfall of an Autocrat”) by R. Kapuściński, “Dzienniki gwiazdowe” (“The Star Diaries”) by S. Lem, “Ferdydurke” by W. Gombrowicz, the epic “Pan Tadeusz” (“Sir Thaddeus”) by A. Mickiewicz, the only Polish translation of “Ulysses” done by M. Słomczyński, 35 poems by C. Miłosz, and 99 poems by W. Szymborska.

All the texts were filtered in order to remove some of the punctuation marks (all except the ones that can functionally end sentences: the periods, the colons and semicolons, the question and exclamation marks) as well as the non-word sequences like numbers. The so-preprocessed texts were subject to further analysis.

3. Results

Although language and language samples are traditionally subject to purely qualitative analysis in the fields of humanities, the language samples consist of symbolic sequences, which can be easily subject to quantitative analysis. Historically, the beginning of quantitative analysis of natural language is usually associated with the name of G.K. Zipf, who was the first to carry out an extensive study of word frequencies in written texts in a few different languages^{17,18}, despite that in fact he also had known predecessors like J.-B. Estoup¹⁹ and E.L. Thorndike²⁰ who did some research in the same direction but far less extensive than the Zipf’s and without any significant impact on science. The main result attributed to Zipf is his eponymous law stating that number F of occurrences of words ordered according to their relative frequency in text samples is, roughly, inverse proportional to their rank R . For the English language, it is:

$$F(R) = \frac{A}{R^\alpha}, \quad \alpha \approx 1. \quad (1)$$

while for other languages α can also be slightly smaller or larger than 1. A stands here for an empirical proportionality constant equal to $0.1T$, where T is the total number of words in a sample (a sample’s length). For single texts samples like books, the above power-law relation is usually well preserved for medium ranks (e.g., $10 < R < 1000$) in the majority of cases, but for the lowest and the highest ranks it breaks down leading to flattening of $F(R)$ for the most frequent words and to its faster decline for the least frequent ones. As regards the multi-piece unions of text samples (corpora), the scaling (1) holds up to the ranks of a few thousand and above them another scaling region with a larger value of α appears^{21,22}. The latter phenomenon can be interpreted as a division of the vocabulary into two sections: the basic vocabulary containing words that are shared by almost all the text samples, thus common to all people, and the specialist vocabulary that can be subject-specific and author-specific²¹. An interesting property of the Zipf law is its similarity to a number of other relations that can be found in different and

4 Authors' Names

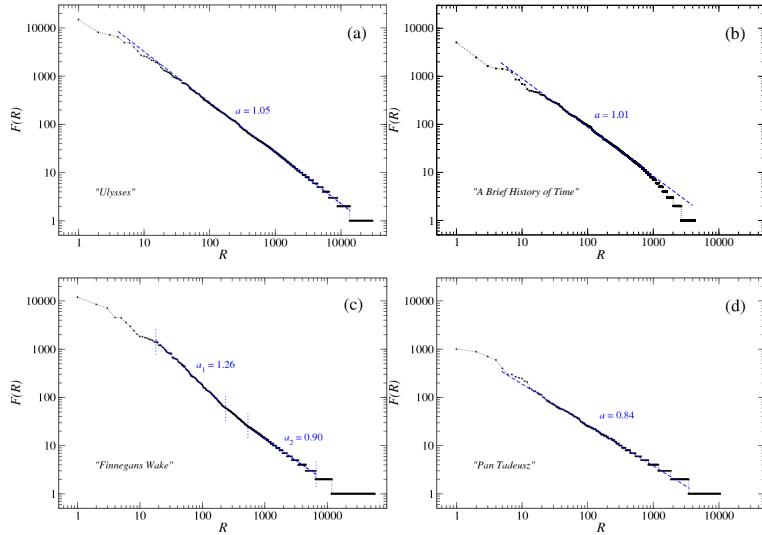


Fig. 1. Rank-frequency distribution of words $F(R)$ in the English texts of (a) “Ulysses” by J. Joyce, (b) “A Brief History of Time” by S. Hawking, (c) “Finnegans Wake” by J. Joyce, and (d) in the Polish text of “Pan Tadeusz” (“Sir Thaddeus”) by A. Mickiewicz. The empirical distributions are compared with the corresponding best fits in terms of a power-law function (Eq. (1)) with the scaling exponent α .

sometimes very distant fields: city population, scientific paper citations, earthquake magnitude, and many more²³.

As regards the rank-frequency relation for text samples, Fig. 1(a) shows such a plot for “Ulysses”. It is notable for its uniquely broad range of ranks (3 decades) for which a power-law scaling holds, which is extremely rarely equalled by other pieces of texts. For example, a similar plot for “A Brief History of Time” in Fig. 1(b) reveals scaling valid for only 2 decades. This means that the vocabulary volume of this book is smaller than it would be expected from the power-law relation holding over all the ranks. On contrary, “Finnegans Wake” (Fig. 1(c)) possesses extremely diverse vocabulary and the rarest words are overrepresented leading to breaking of the Zipf-like relation in the opposite direction as compared to the Hawking’s book. Such a situation does not surprise us, however, since “Finnegans Wake” is known to be a highly experimental piece of text comprising words from many languages. Next, Fig. 1(d) shows a rank-frequency plot obtained for the Polish text of “Pan Tadeusz”. A well-fitted power-law function with $\alpha = 0.84$ describes the plot, whose slope is much smaller than for typical English texts and even for typical Polish ones ($\alpha = 0.94$ ^{24,25}), while it is characteristic for the works of A. Mickiewicz²⁶.

Characterization of a text sample by means of the Zipf plot is informative in respect to vocabulary volume of an author and mutual relations between the most common and other words, but it is insensitive to any kind of correlations possibly present in the sample. In order to incorporate correlations into our analysis, we

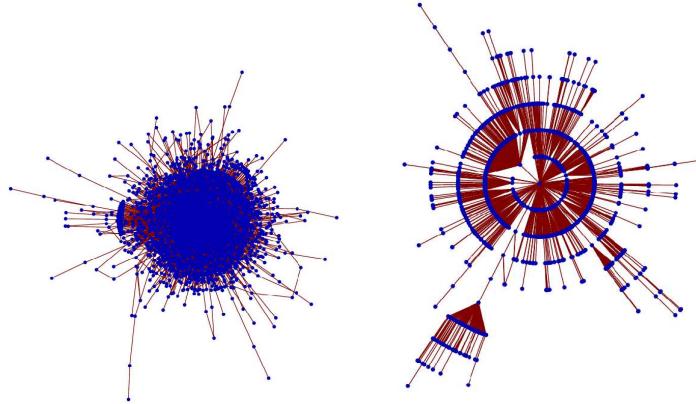


Fig. 2. Binary network representation (left) and minimal spanning tree (right) of an exemplary text sample: “Statistical mechanics of complex networks”. Both pictures correspond to the succession network.

create network representations of each of the text samples studied here. We choose such a representation in which different words are regarded as different network nodes. One type of interesting correlations that can be quantified in this way is the adjacency relation between pairs of words. Two words are considered related and their nodes linked by an edge if they are the nearest neighbours at least once in a sample. For a given word, there are two possible relations with its neighbours: precedence and succession. The former is when a neighbour precedes the considered word, while the latter is in the opposite case. In this context, we study two networks: the precedence (left-side neighbourhood) network and the succession (right-side neighbourhood) network. We decided to consider only the neighbours that belong to the same sentence and neglect the inter-sentence adjacency. This does not influence our results, however: a preliminary analysis carried on a few text samples showed that there is no qualitative difference of the results between these cases. This, of course, might be a consequence of a much smaller number of such inter-sentence pairs (roughly, less than 10% of all pairs).

By construction, our networks can be either binary or weighted. In the former case, we consider two nodes to be linked by an edge if the respective words are neighbours at least once in a text but we do not pay attention to how many times they neighbour each other. In contrast, in the latter case, we may count the number of such occurrences and attribute a corresponding weight to each edge. Examples of both situations are presented in Fig. 2, where a binary network (left) and a minimal spanning tree calculated from a weighted network (right) are created for an exemplary piece of text.

To begin with, let us calculate a cumulative distribution $P(X \geq k)$ of node degrees k for different texts. This is one of the most informative quantities since it

6 Authors' Names

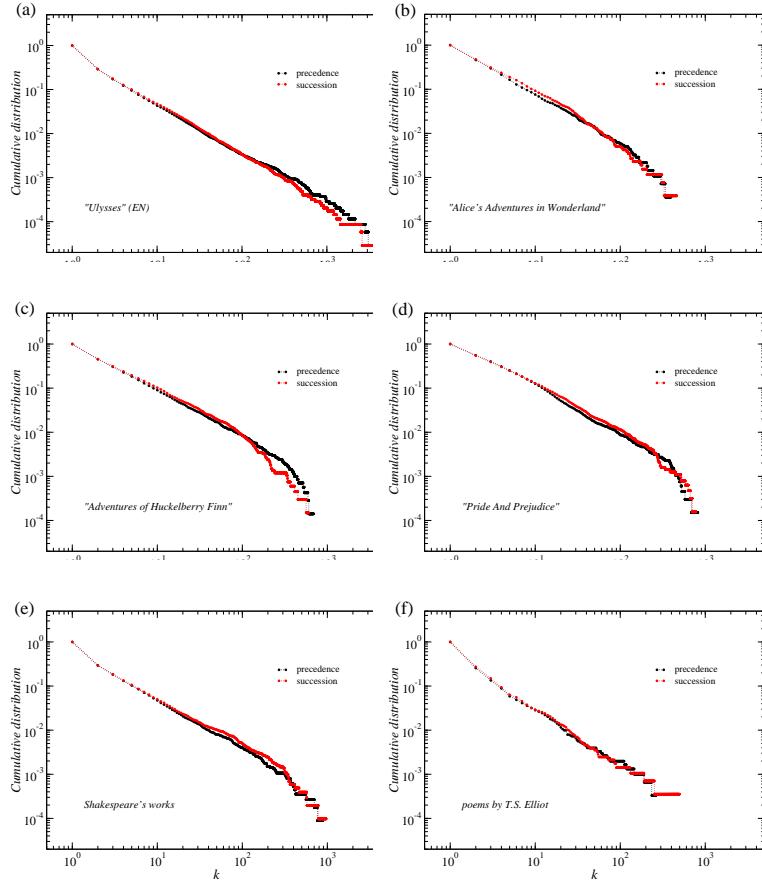


Fig. 3. Cumulative distributions $P(X \geq k)$ of the node degrees k for the word-adjacency network representations of English literary texts: (a) “Ulysses”, (b) “Alice’s Adventures in Wonderland”, (c) “Adventures of Huckleberry Finn”, (d) “Pride and Prejudice”, (e) 4 Shakespeare dramas, and (f) 25 poems by T.S. Elliott. The precedence and succession networks are shown simultaneously in each panel.

allows one to detect a hierarchical and scale-free structure of a given network²⁷. Since such distributions for weighted networks carry basically the same information as the Zipf plots, we restrict our analysis to the binary networks only. Fig. 3 exhibits cumulative distributions $P(X \geq k)$ for selected literary texts in English. Interestingly, although there are clear differences between the distributions for different texts, all the texts studied (including other not shown here) reveal the scale-free or almost scale-free dependence for some range of k , with the scaling exponents $1 < \beta < 2$ being in agreement with the results from other studies of different systems^{28,29}. It happens for some texts that the precedence and the succession networks visibly differ from each other. We do not inspect this issue in more detail, but a source of this difference might be either author-specific or text-specific.

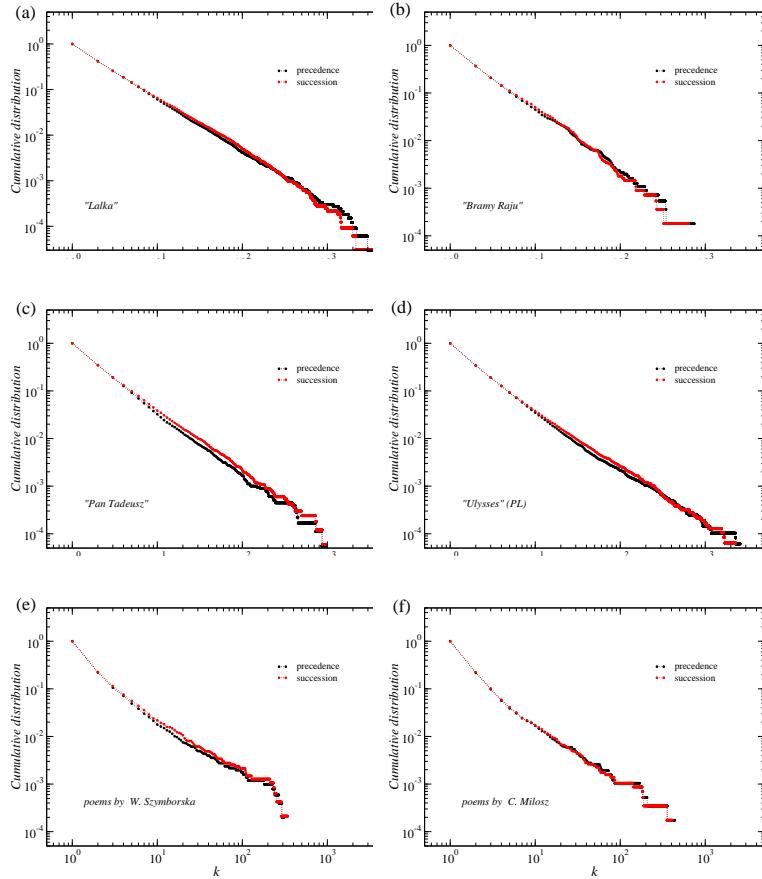


Fig. 4. Cumulative distributions $P(X \geq k)$ of the node degrees k for the word-adjacency network representations of Polish literary texts: (a) “*Lalka*”, (b) “*Bramy Raju*”, (c) “*Pan Tadeusz*”, (d) a Polish translation of “*Ulysses*”, (e) 99 poems by W. Szymborska, and (f) 35 poems by C. Milosz. The precedence and succession networks are shown simultaneously in each panel.

In Fig. 4, the $P(X \geq k)$ distributions are shown for selected Polish literary texts. For prose ((a)-(d)), the scale-free slopes of these distributions are even better visible than for the English texts in Fig. 3. The same refers to poems except the overrepresentation of nodes with small k in the case of Polish poetry (Fig. 4(e)-(f)). This overrepresentation probably stems for the fact that poetry, which needs a specific rhythm, imposes strong restrictions on the words that can be used in particular places.

In order to compare statistical properties of node degrees for different types of texts and the two languages, from the texts considered in this work, we create five separate corpora containing English prose, English poetry, English scientific texts, Polish prose, and Polish poetry. Then for each corpus, we calculate a node degree cumulative distribution $P(X \geq k)$. Fig. 5 shows these distributions along with their

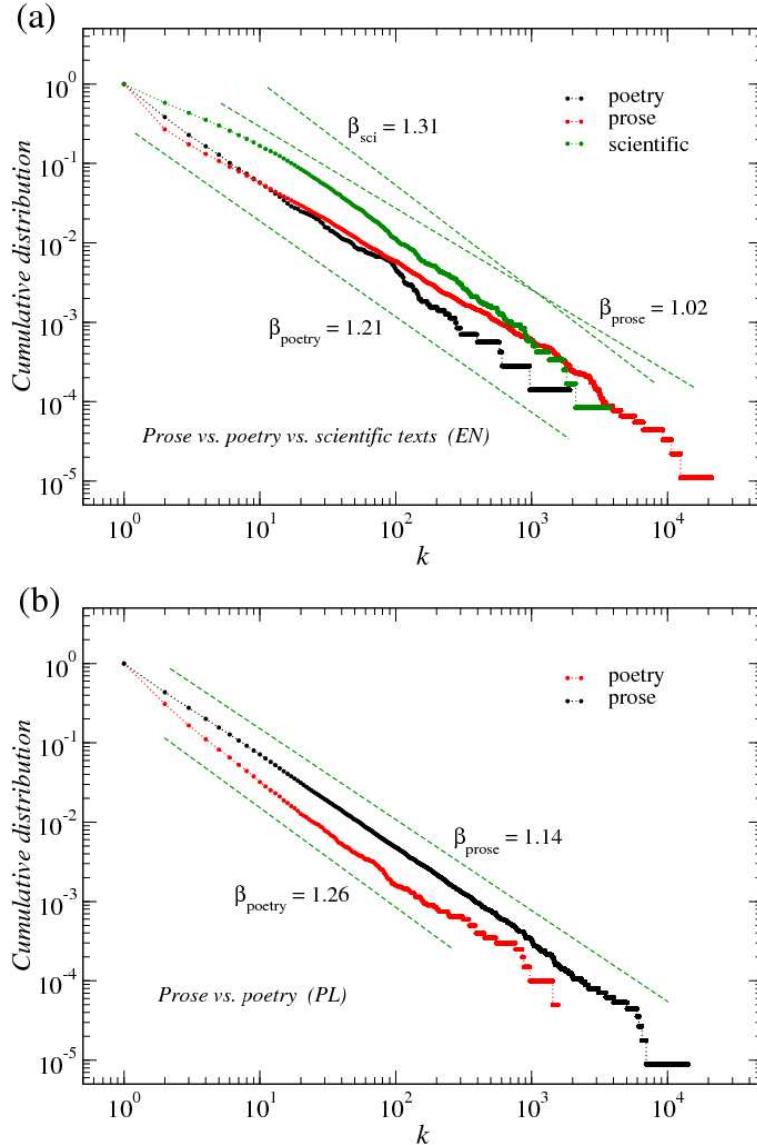


Fig. 5. Cumulative distributions $P(X \geq k)$ of the node degrees k for the word-adjacency network representations of English (a) and Polish (b) corpora containing prose, poetry or scientific texts. Slopes of the best-fitted power laws are indicated by the dashed lines and the values of the scaling exponent β .

power-law slopes. Regarding prose, the distribution for the English language has smaller slope than its counterpart for the Polish language. However, the picture for poetry looks different: the node degree distribution is steeper in the case of English. Roughly, as regards the corpora of Polish prose and of Polish poetry, the distribu-

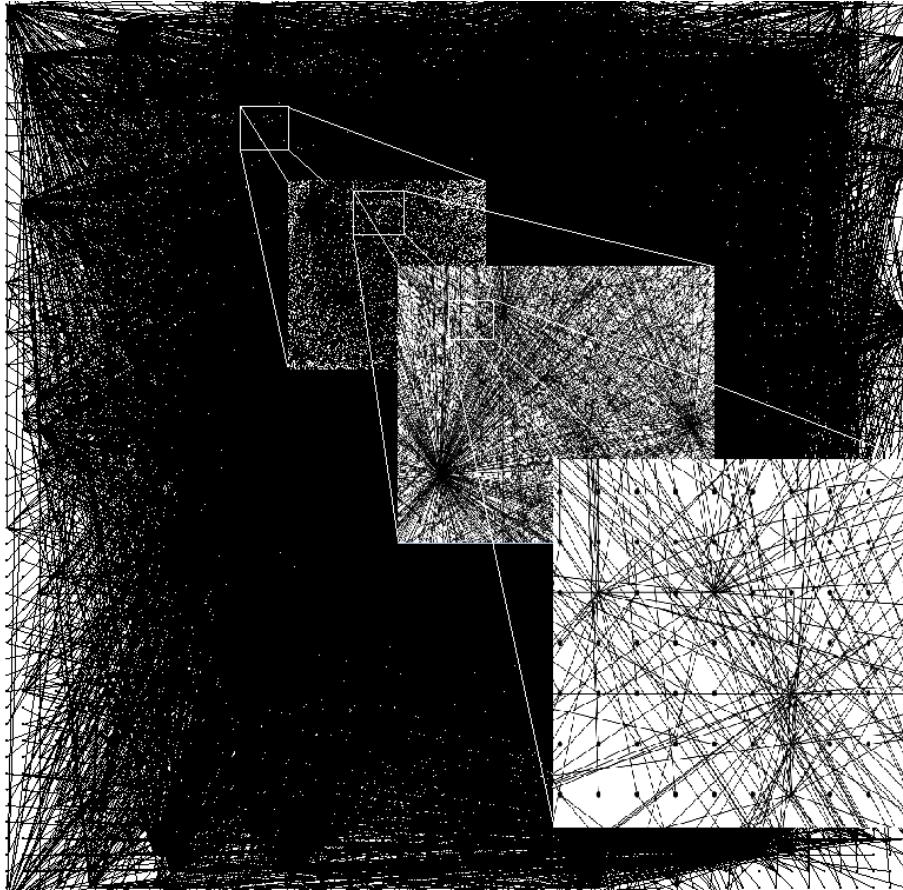


Fig. 6. Visualization of a network representation of an exemplary text sample showing a hierarchy of nodes.

tions look similar, which is not the case for their English counterparts. $P(X \geq k)$ for the corpus of scientific texts written in English reveals the most steep slope with $\beta = 1.31$. This is not surprising, however, since many scientific texts are full of mathematics and related formal names and expressions, which make the vocabulary poorer than in the case of literary works, which do not have any vocabulary restrictions (see Fig. 5(a)). As regards $P(X \geq k)$ for the individual scientific papers, some of them do not reveal any trace of scaling while other are clearly scale-free. This strongly depends on the relative amounts of standard description and strict mathematical language: the less mathematics is there, the better scaling can be observed.

Our results for both languages indicate that the adjacency networks in both representations are strongly non-democratic with a clear hierarchy of hubs. Indeed, Fig. 6 confirms this conclusion by showing both global and local hubs with large

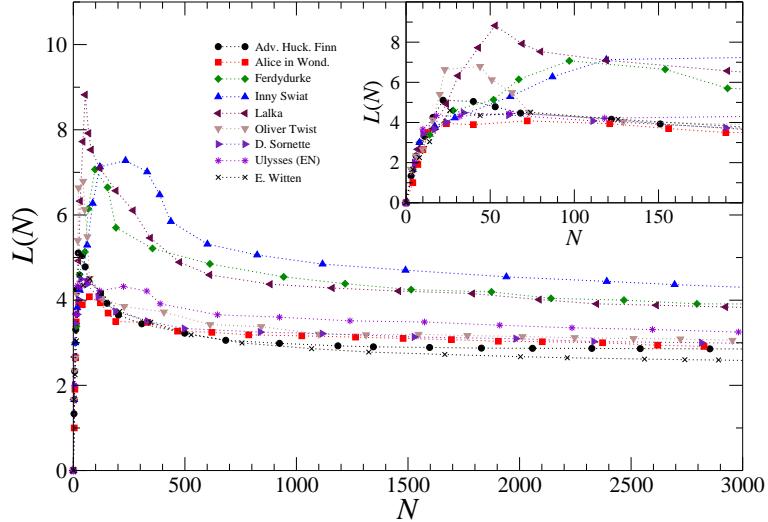


Fig. 7. The average shortest path length L as a function of the number of nodes N for exemplary texts considered in this work (only $N < 3000$ is shown because the texts differ in their vocabulary volume). The inset shows magnification of the small- N region.

values of k surrounded by clouds of peripheral nodes with $k \approx 1$. Other topological properties of the networks can be characterized by their spatial extension and inclination of nodes to form clusters. The former can be quantitatively described by the average characteristic path length L expressing the average node-to-node distance. For a binary network it is defined by:

$$L = \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{N}, i \neq j} L_{ij}, \quad (2)$$

where $L_{i,j}$ is the length of the shortest path connecting the nodes i and j (the length of each path is defined as a number of edges this path passes through). Magnitude of L and, especially, its dependence on N is different for different network types. It typically grows fast for regular lattices and chains ($L \sim N$), moderately fast for random networks of the Erdős-Rényi type ($L \sim \ln N$), the small-world networks ($L \lesssim \ln N$) and for the Barabási-Albert networks ($L \sim \ln N / \ln \ln N$)¹⁶, slowly for the ultrasmall networks ($L \sim \ln \ln N$)³⁰, while for densely connected networks it can roughly be independent of N . For our networks, values of this quantity (for the complete texts) belong to the interval: $2.7 \leq L \leq 3.8$, which generally falls into the small-world networks regime. However, the asymptotic behavior of $L(N)$ is significantly different from the small-world one since for $N \gg 1$, L is decreasing function of N (Figure 7). This is because, for large N , the vocabulary volume V (here, $V = N$) used for writing a text grows much more slowly than the length T of the text (which is expressed by a general relation $V \sim T^\delta$, where $0.4 \lesssim \delta \lesssim 0.6$ - the so-called Heaps law) and this leads to increasing density of edges.

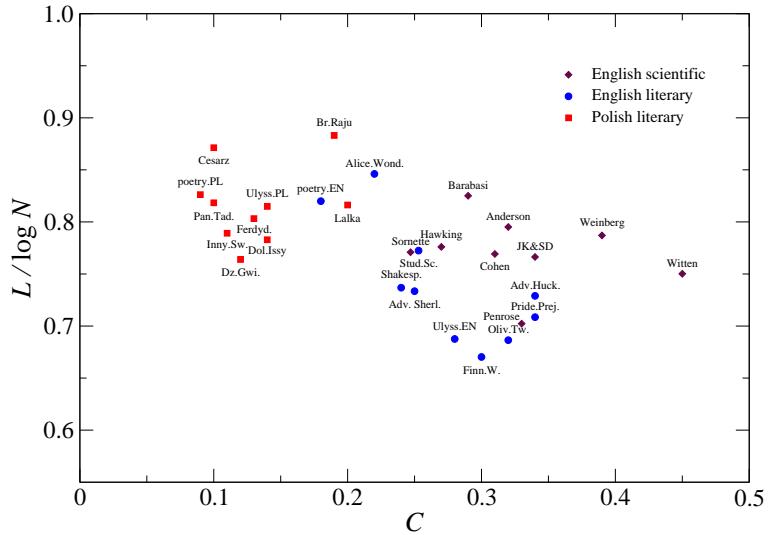


Fig. 8. The clustering coefficient C vs. the average shortest path length $L/\log N$ for the texts considered in this work. Different groups of texts are denoted by different symbols.

The clustering coefficient C for an undirected binary network is expressed by:

$$C = \frac{1}{N} \sum_i \frac{\sum_{j,m} a_{ij} a_{jm} a_{mi}}{k(k-1)}, \quad (3)$$

where a_{pq} are binary edges (equal to 1 for the existing edge and 0 otherwise). Its value is typically small for the Erdős-Rényi networks ($C \sim N^{-1}$) and for the Barabási-Albert networks ($C \sim N^{-0.75}$)¹⁶, while it is large (and independent of N) for the small-world networks of the Watts-Strogatz type³¹.

For all the texts except the highly mathematical papers by E. Witten and S. Weinberg, values of C are in the interval $0.09 \leq C \leq 0.34$, but unlike the shortest path length, here we observe rather clear separation between Polish and English novels (including the Shakespeare's drama): $0.10 \leq C \leq 0.20$ for the Polish ones and $0.22 \leq C \leq 0.34$ for the English ones. The poetry in both languages tends to have smaller clustering coefficient than any other considered piece of text: $C_{PL} = 0.09$ and $C_{EN} = 0.18$. These results can be seen in a scatter plot in Figure 8, which shows values of the clustering coefficient and the shortest path length for all the texts considered in this work. Indeed, the Polish and the English texts occupy different regions of the (C,L) plane with the Polish ones being characterized by smaller C . For English, one can also see that scientific texts may have different properties than the literary texts, especially if they contain much mathematics (Witten, Weinberg). If they are more descriptive than mathematical, their properties can resemble the properties of literary texts (Sornette, Hawking, Penrose).

It is interesting to note that unlike other network types mentioned above, for our adjacency networks, C is an increasing function of N and, typically, its dependence

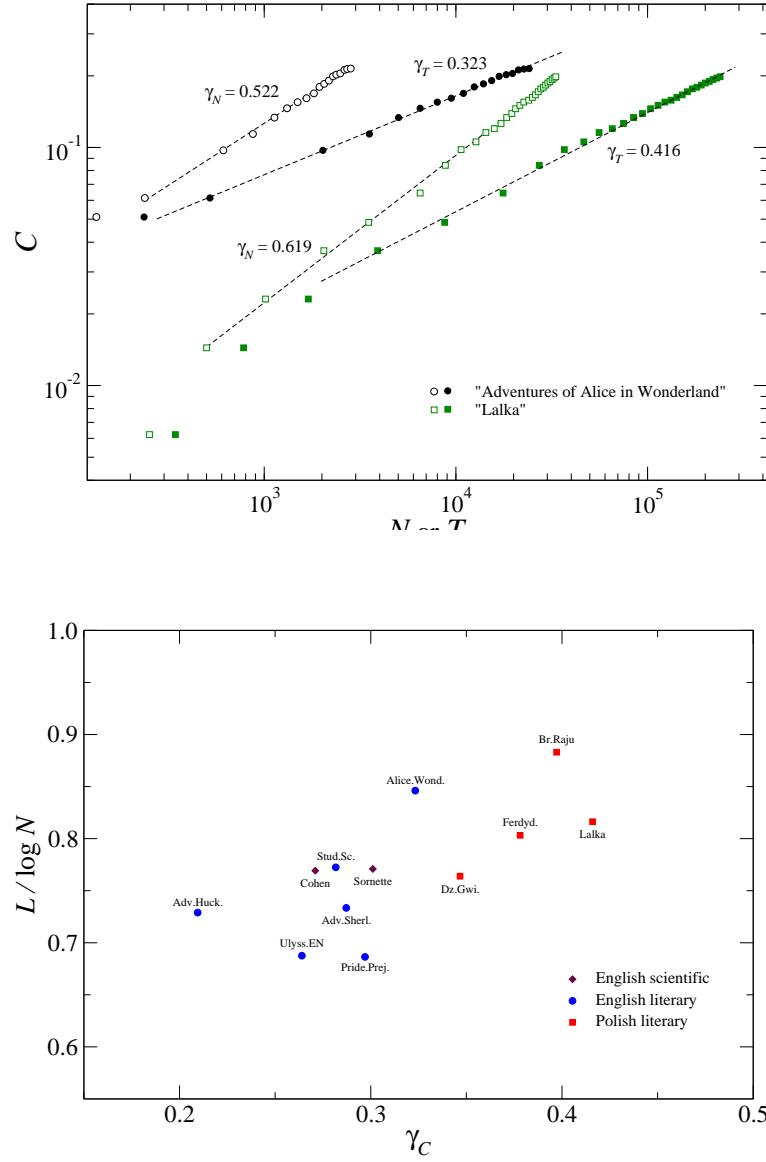


Fig. 9. (a) The clustering coefficient C dependence on the number of nodes N and the text length T for two exemplary texts with the observed scaling behavior $C \sim N^{\gamma_N}$ or $C \sim T^{\gamma_T}$. The scaling indices γ_N and γ_T are shown together with the corresponding least-squares fits. (b) The scaling index γ_T vs. the average shortest path length L for those texts for which $C(T)$ was at least partially power-law. Different groups of texts are denoted by different symbols.

is either power-law $C \sim N^{\gamma_N}$ (at least for large values of N) with the scaling index $\gamma_N < 1$ or not far from power-law. A similar power-law behavior can be seen for $C(T)$, where T is the text length, but this is not surprising due to the

already-mentioned Heaps law: $N \sim T^\delta$. Figure 9a shows two examples of such power-law behavior for one English and one Polish text. As it was the case with C , the English and the Polish texts have distinct values of γ_T , with the latter being significantly larger. Calculated values of this index for the texts that show clear power-law dependence of $C(T)$ are collected in Figure 9b. Although γ_T cannot be estimated for all the texts from Figure 8, the separation of the languages is clear also here.

The observed behavior of $L(N)$ and $C(N)$ suggests that the networks considered in this work have their own specific structure that clearly distinguishes them from the most well-known network structures. They cannot be counted as small-world networks even though the average characteristic path length is relatively short. They also differ from the Barabási-Albert networks despite the fact that some of them show the scale-free structure and from the random (Erdős-Rényi) ones.

4. Conclusions

We presented several results from our quantitative study of statistical and network properties of literary and scientific texts written in two substantially different languages: English and Polish. We transformed the text samples into word-adjacency networks defined by the nodes representing individual words and the edges representing pairs of directly neighbouring words. For the majority of the studied literary texts in both languages, the corresponding networks revealed the scale-free structure, while this was rarely the case for the scientific texts. We also showed that there are differences in node degree distributions between prose and poetry, especially in English. Poetry has a detectable steeper distribution's slope than has prose. The slope for scientific texts is even steeper than for poetry, but this can be explained by typically poorer vocabulary in the former case. Despite these differences, all the network representations of texts were hierarchical with a few important hubs and the majority of less important nodes. No qualitative and quantitative difference between the languages was noticed in this respect. This picture changed completely if we looked at other network statistics like the clustering coefficient and the average shortest path length. The English texts appear to possess more clustered structure, while the Polish ones were less clustered. This result was attributed to differences in grammar of both languages, which was also indicated in the Zipf plots. Our results suggest that the word-adjacency networks cannot fully be described by any of the Erdős-Rényi, the Watts-Strogatz, and the Barabási-Albert models even though these networks exhibit certain characteristics of the latter two models. Such networks will be a subject of our forthcoming study.

References

References

1. M.H. Christiansen, N. Chatera, Language as shaped by the brain, *Behav. Brain Sci.* **31**, 489 (2008)

14 *Authors' Names*

2. M.A. Nowak, D.C. Krakauer, The evolution of language, Proc. Natl. Acad. Sci. USA **96**, 8028 (1999)
3. M.A. Nowak, J.B. Plotkin, V.A.A. Jansen, The evolution of syntactic communication, Nature **404**, 495 (2000)
4. P.W. Anderson, More is different, Science **177**, 393 (1972)
5. J. Kwapień, S. Drożdż, Physical approach to complex systems, Physics Reports 515, 115-226 (2012)
6. A.P. Masucci, G.J. Rodgers, Network properties of written human language, Phys. Rev. E **74**, 026102 (2006)
7. S.M.G. Caldeira, T.C. Petit Lobão, R.F.S. Andrade, A. Neme, J.G.V. Miranda, The network of concepts in written texts, Eur. Phys. J. **49**, 523 (2006)
8. L. da Fontoura Costa, O.N. Oliveira Jr., G. Travieso, F. Aparecido Rodrigues, P.R. Villas Boas, L. Antiquera, M. Pahlares Viana, L.E. Correa da Rocha, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Adv. Phys. **60**, 329 (2011)
9. S. Zhou, G. Hu, Z. Zhang, J. Guan, An empirical study of Chinese language networks, Physica A **387**, 3039 (2008)
10. R.M. Roxas-Villanueva, M.K. Nambatak, G. Tapang, Characterizing English poetic style using complex networks, Int. J. Mod. Phys. C **23**, 1250009 (2012)
11. E. Witten, (1) Analytic continuation of Liouville theory; (2) Khovanov homology and gauge theory; (3) Knot invariants from four-dimensional gauge theory; (4) Fivebranes and knots; (5) A new look at the path integral of quantum mechanics; (6) An algebraic construction of boundary quantum field theory; (7) The omega deformation, branes, integrability, and Liouville theory; (8) Analytic continuation of Chern-Simons theory; (9) Geometric langlands and the equations of Nahm and Bogomolny; (10) Geometric langlands from six dimensions; all papers available from <http://www.arxiv.org>
12. E.D.G. Cohen, (1) Jet-like tunneling from a trapped vortex; (2) Stick-slip motion of solids with dry friction subject to random vibrations and an external field; (3) Path integral approach to random motion with nonlinear friction; (4) Steady state work fluctuations of a dragged particle under external and thermal noise; (5) Anomalous fluctuation relations; (6) Fluctuation properties of an effective nonlinear system subject to Poisson noise; (7) Entropy, probability and dynamics; (8) Properties of nonequilibrium steady states: a path integral approach; (9) Gaussian approximation to single particle correlations at and below the picosecond scale for Lennard-Jones and nanoparticle fluids; (10) Nonequilibrium steady state thermodynamics and fluctuations for stochastic systems; all papers available from <http://www.arxiv.org>
13. S. Weinberg, (1) Quantum contributions to cosmological correlations; (2) Effective field theory for inflation; (3) A tree theorem for inflation; (4) Non-Gaussian correlations outside the horizon II: the general case; (5) Living with Infinites; (6) Effective field theory, past and future; (7) Asymptotically safe inflation; (8) Six-dimensional methods for four-dimensional conformal field theories; (9) Pions in large-N quantum chromodynamics; (10) Collapse of the state vector; all papers available from <http://www.arxiv.org>
14. P.W. Anderson, (1) Simple explanation of Fermi arcs in cuprate pseudogaps: a motional narrowing phenomenon; (2) A Fermi sea of heavy electrons (a Kondo lattice) is never a Fermi liquid; (3) A Gross-Pitaevskii treatment for supersolid He; (4) Incoherent tunneling amplitude in high- T_c cuprates; (5) Transport anomalies of the strange metal: resolution by hidden Fermi liquid theory; (6) Beyond the Fermi liquid paradigm: hidden Fermi liquids; (7) Personal history of my engagement with cuprate superconductivity; (8) Hidden Fermi liquid: self-consistent theory for the normal state

of high- T_c superconductors; (9) The ground state of the Bose-Hubbard model is a supersolid; (10) Theory of supersolidity; all papers available from <http://www.arxiv.org>

- 15. D. Sornette, Critical market crashes, *Phys. Rep.* **378**, 1 (2003)
- 16. R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**, 47 (2002)
- 17. G.K. Zipf, Selective studies and the principle of relative frequency in language, MIT Press (1932)
- 18. G.K. Zipf, Human behavior and the principle of least effort, Addison-Wesley (1949)
- 19. J.-B. Estoup, Gammes sténographiques. Méthodes et exercices pour l'acquisition de la vitesse, Institut Sténographique de France (1916)
- 20. E.L. Thorndike, A teacher's word book of 20,000 words, Teacher's College (1932)
- 21. M.A. Montemurro, Beyond the Zipf-Mandelbrot law in quantitative linguistics, *Physica A* **300**, 567 (2001)
- 22. R. Ferrer i Cancho, R.V. Solé, Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited, *J. Quant. Ling.* **8**, 165 (2001)
- 23. M.E.J. Newman, Power laws, Pareto distributions and Zipf's law, *Contemp. Phys.* **46**, 323 (2005)
- 24. J. Kwapienie, S. Drożdż, A. Orczyk, Approaching the linguistic complexity, Lecture Notes of the Institute for Computer Science, Social Informatics and Telecommunication Engineering **4**, 1044 (2009)
- 25. J. Kwapienie, S. Drożdż, A. Orczyk, Linguistic complexity: English vs. Polish, text vs. corpus, *Acta Phys. Pol. A* **117**, 716 (2010)
- 26. A. Orczyk, Zjawisko złożoności a struktura języka polskiego: analiza leksykologiczna wybranych źródeł literackich i pozaliterackich narzędzi statystycznej, Master thesis (in Polish), WFiS AGH (2008)
- 27. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: structure and dynamics, *Phys. Rep.* **424**, 175 (2006)
- 28. S. Redner, How popular is your paper? An empirical study of the citation distribution, *Eur. Phys. J. B* **4**, 131 (1998)
- 29. A.-L. Barabási, R. Albert, Emergence of scaling in random networks, *Science* **286**, 509 (1999)
- 30. R. Cohen, S. Havlin, Scale-free networks are ultrasmall, *Phys. Rev. Lett.* **90**, 058701 (2003)
- 31. D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393**, 440 (1998)